大語彙を対象とした音声対話インタフェースにおける自然な応答生成

大森 久美子[†,††] 斎藤 博昭[††]

本稿は,「思い込み応答」戦略を取り入れた大語彙音声対話インタフェースを提案する．この戦略は，人間同士の対話において発話対象が広範囲に及ぶ場合，聞き間違えにくい対象と間違えやすい対象が存在することに着目したもので，聞き間違えやすい対象を誤認識しても利用者にストレスを与えないことを利用している．大語彙として16万種の個人姓に焦点を当て，音声認識精度と語彙網羅率の観点から，聞き間違えてはならない10,000種の思い込み対象を選択できた．更に，思い込みが外れた場合への対応として，思い込みの結果を利用者に応答として提示している時間を利用して，思い込み範囲外の残りの姓を対象とした裏認識処理を並行して進める仕組みを提案した．市販の認識エンジンを利用して，この仕組みと思い込み応答を組み合わせた個人姓確定インタフェースを実装した．思い込み応答は，現状の音声認識技術を用いたインタフェースにおいて，入力対象が大語彙であってもストレスを与えない結果を利用者に提示できる戦略であることを確認した．

キーワード: 音声対話インタフェース，対話制御方式，大語彙，住所確定タスク，姓名確定タスク，利用者満足度

A Spoken Dialogue Interface through Natural and Efficient Responses

KUMIKO OHMORI[†,††] and HIROAKI SAITO[††]

This paper proposes a new dialogue control method with "presuppositional responses" to realize a large number of target words towards an efficient spoken dialogue interface. This strategy comes from human characteristics in that people tend to presuppose the utterance to be familiar or frequently-spoken. The strategy is verified through huge data of human recognition of 160,000 sir names. We introduce heuristics to determine what words are to be presuppositional; presuppositional words should cover as many frequently-used ones as possible, while they should be small for high-accurate speech recognition. We report a successful implementation of a dialogue interface using a conventional speech recognition device. We resolve the situations when speech recognition fails or when the corrent answer is not included in presuppositional words in order not to irritate the user with unnecessary or detoured questions. Realtime and natural responses are attained through parallel search of non-frequent words as well as presuppositional ones.

KeyWords: *Speech dialogue interfaces, Dialogue control methods, A large number of target words, address retrieval, name retrieval, customer satisfaction*

† (株) NTT データ技術開発本部, NTT DATA CORPORATION Research and Development Headquarters
尚，この研究は NTT 情報流通プラットフォーム研究所在籍中に行ったものである．
†† 慶應義塾大学大学院理工学研究科開放環境科学専攻, Graduate School of Science and Technology, Keio University

1　はじめに

近年，音声認識技術や言語処理技術，計算機の処理能力の向上により，情報検索をはじめとする各種タスクを音声認識を介して実現する音声対話インタフェースへの期待が急速に高まっている (Nielsen and Baekgaard 1992; Godden,Brill,Glass,Pao,Phillips,Porifroni,Sneff and Zue 1994; Zue,Seneff,Glass,Goddeau,Goddin,Pao,Phillips and Porifroni 1994; Zeigler and Mazor 1995; Godden,Meng,Polifromi,Seneff and Busayapongchai 1996; Ferguson and Allen 1998; Nakano,Dohsaka,Miyazaki,Hirasawa,Tamoto,Kawamori,Sugiyama and Kawabata 1999)．同時に，音声対話インタフェース実現のための対話制御方式も数多く提案されている (Niimi,Takigawa and Nishimoto 1995; 新美，小林 1995; Niimi,Nishimoto and Kobayashi 1997; 新美 1998; 菊地，白井 2000; Chu-Carroll 2000)．音声による入力は，操作に熟練を必要としないため利用者にとっては使い勝手が良く，入力速度はキー入力に比べ3〜4倍，手書き文字入力に比べ8〜10倍速いと言われている (古井 1998)．更に，他の器官を同時に使っての並行作業が可能であるという利点を有する．また，サービス提供者にとってはオペレータコストの削減に繋がる．

実用サービスのフロントエンドとして音声認識を適用するためには，不特定多数の話者の入力に対して，迅速かつ正確に応答する必要がある．音声認識の性能は，発話様式によって大きな影響を受けることが指摘されている (村上，嵯峨山 1991)．最も単純なシステム主導，一問一答形式の単語認識でも，対象単語数が増えるほど誤認識は避けられず処理時間を要する．更に，音声認識は利用される環境や発話状況により誤認識を生じる場合も多く，公衆電話網は帯域が狭いため認識精度が落ちる．

我々は，顧客が入力する住所や姓名の確定をタスクとする音声対話インタフェースの実現に向け，検討を進めている．音声認識技術においてエンジンの出力結果が正しいか否かを判断するには，発話者本人に正誤を確認するしか方法はない．特に，不特定多数の話者が入力する住所や姓名などの大語彙を認識対象とする場合，正確な応答を返すことは困難である．音声対話インタフェースの現状は，(1) 個々の質問において利用者が予期しない対象への誤認識が多い (2) 正誤確認と誤認識を修正するための再入力要求が繰り返される，という2つの要因から利用者満足度が獲得できていない．従って，音声対話インタフェースの実用化のためには，上記2つの要因解決が必須となる．

本稿は，上記要因 (1) の解決に焦点を当て，人間が発話を聞き取る際の傾向に着目し「思い込み応答」という聞き取り結果の確認手法を提案する．そして思い込みによる認識結果の確認が，入力対象が大語彙であっても利用者にストレスを与えないことを検証する．この思い込み応答は音声入力の応答に特化したものではないが，本稿では音声入力を例として以下議論を進める．その他への適用については6章の今後の課題で述べる．

以下2章では，大語彙を対象とした音声対話インタフェースの現状の課題について述べる．3章では人間の対話における思い込み戦略を検証し，4章では，市販の認識エンジンを用いて思

い込み対象の選択方法について分析する．5章では，思い込み戦略を取り入れた聞き取り確認手法を提案し，実装及び評価を通してその有効性を検証する．最後に6章にて，まとめ及び今後の課題について述べる．

2　大語彙音声対話インタフェースの課題

2.1　現行の音声対話インタフェース

音声認識技術の限界から，音声対話インタフェースは天気案内や株価照会，星占いなど対象語数の限られた分野でしか実現されていない[1]．特に，住所や姓名の確定をタスクとした音声対話インタフェースは，コールセンターの業務効率化などに有効であることから提案も多い (赤堀，加藤，北岡 1995; 荒井，吉岡，嵯峨山，山田，野田，井本，菅村 2000; 吉岡，荒井，菅村，嵯峨山 1997)．しかし対象が大語彙であるため，最終的に確定したい住所や姓名を最初の入力対象に設定するのは困難であり，高精度な結果が期待できる小語彙を入力対象とした質問を組合せることで対象を絞り込み，確定を実現している．

住所確定に関しては，ボイスポータルの天気案内サービスに代表されるように，都道府県，市区郡，大字という階層的なデータ構造を利用して，上位から順に確定を行う提案が多い[2]．これは，情報検索におけるディレクトリ検索方式を音声入力に適用したものである．現在，日本全国には47都道府県，4,100市区郡，173,600の大字の約18万種の地名が存在する[3](自治省 1998)．現状の音声認識技術を利用して，18万種の地名を一度に認識対象とすることは非現実的である．上位層から順に対象を絞り込みながら確定することで，各質問毎の応答は，リアルタイム性及び利用者にストレスを与えない精度を持ったものとなっている (亀田，藤崎 1997)．

音声入力型ディレクトリ検索方法には，以下の課題がある．

(1)　利用者の入力対象を階層化して一度の質問における認識対象数を減少させ，上位から順に入力を確定する．そのため，対象の階層数分の〈入力要求，正誤確認〉が必要になる．

(2)　上位階層が確定しないと下位の対象を絞り込むことができない．そのため，各階層において認識結果が正解であるという確認が得られるまで，下位階層の入力要求へ進むことができない．

(1) に関して，堂坂らは，個人名や部署名など予め規定したスロットを順に埋めていくタスクにおいて，現在のスロットとその値からスロット間の依存関係を考慮して，正誤確認の回数を最小化する方法を提案している (堂坂，安田，相川 2002)．しかし，質問順序は対象語数に制

1 http://www.nomura.co.jp/service/kabukadial.html, http://www.ufj-tsubasa.co.jp/contact/index.html など.

2 http://www.ntt.com/v-portal/, http://www.voizi.net, http://www.jmscom.co.jp/automat/index.html など.

3 中央区，本町など，全国に同一地名が複数存在する場合は1件にカウント.

限され，システム主導に予め決められる．これに対して伊藤らは，効率を優先した質問順序がかえって機械特有の不自然さに繋がり，これが音声対話インタフェースが実用レベルで受け入れられない理由であると考察している (伊藤，駒谷，河原 2002)．

一方，個人姓名を入力対象としたインタフェースはパソコンのサポートセンターなどでの実用化例がみられる[4]．しかし階層的データ構造を持たない姓名には，ディレクトリ検索方式を適用できないため，数百名程度の事前登録会員のみに対象を限定するなど，対象を小規模にしなければ実現は困難である．夜間や休日などオペレータの業務時間外に自動応答で対応するサービスでは，音声認識技術を利用して，利用者の入力を正確に認識し確定することが困難なため，利用者に姓名や電話番号を留守番電話に録音してもらい，後日人間が聞いてコールバックする方法が一般的である[5]．新規加入申込みなど，予め対象が限定できない姓名を対象とする実用インタフェースでは，実在者数の偏りを利用して実在頻度順位上位から数千種の頻出姓名のみを対象としている場合も多い．しかしこれでは希少姓を入力する利用者に対応できない．現状，多数の利用者からのアクセスが予想されるサービスでの姓名確定の実用化例は存在しない．

2.2　大語彙音声対話インタフェース実現に向けて

我々は，資料や商品配送先の特定やチケット予約などにおいて，申込者の特定を行うための住所や姓名の確定をタスクとした音声対話インタフェースの実現を目指す．第一段階として，音声認識技術及びそれを利用したインタフェースの現状を踏まえ，利用者の入力対象が大語彙である場合，利用者にストレスを与えない応答を提供する必要があると考える．

利用者から住所，姓名，年齢などを聞き取り，商品カタログ送付を専門業務とするコールセンターへの1日のアクセスを分析した[6]．オペレータから住所を尋ねられると，首都圏，及び各都道府県の県庁所在地に在住の利用者は，最初に港区，横浜市，名古屋市など住所の市区郡名を答える場合が多く，中でも東京23区内在住の利用者は，六本木，虎ノ門のように字名を答える傾向が多く見受けられた．それ以外の利用者は，都道府県名を最初に発声する場合が多い．この分析から，住所確定に関しては，どのレベルの地名が入力されても対応可能であることが望ましい．また，姓名に関しては，姓16万種，名は男性だけでも8万種存在する (星野，加藤，永田 1993)．現状の認識技術を利用して，この大語彙を一度に認識対象とすることは非現実的である．

4 http://vcl.vaio.sony.co.jp/info/ivr/index.html, http://support.jp.dell.com/jp/jp/spm/phone/など．
5 http://www.ntt.com/shop/toiawase/email.wbt
6 所在地は東京都港区，1日の平均アクセス数は5,000コール．

3　思い込み戦略

本章は，人間同士の対話における聞き取り傾向を分析し，話し手の発話対象が大語彙に及ぶ場合の聞き取り戦略を提案する．

3.1　人間の対話における思い込みの検証

我々は，人間同士の対話においても聞き間違えがあることに着目した．すなわち，人間は初対面の相手に住所や名前を尋ねる際，聞き覚えのある地名や姓名は正しく聞き取れるが，初めて耳にするものに対しては正しく聞き取れたかどうか確信を持てない場合が多い．我々は，聞き手は話し手の発話対象を全て網羅しているわけではなく，過去の経験などから発話対象に対する思い込みが働き，その思い込んだ範囲内から聞き取った内容を探し出そうとするという仮説の検証を進めた．聞き手は思い込みのために，予測した範囲外の対象が発話された場合に聞き間違えを起こすと予想した．本節は，人間の対話における思い込み戦略を実証するために，約16万種の日本人姓を対象とした聞き取り試験を実施した．聞き手は日頃から良く耳にする姓は正しく聞き取れるが，聞き慣れない珍しい姓ほど聞き間違えやすいという結果が得られた．以下，試験概要を示す．

男女合わせた10名の被験者[7]に，電話回線を介して日本人の姓[8]を聞いてもらい，結果の書き取りを依頼した．被験者には事前情報として日本人の姓が発話されることのみを伝え，1件につき1度の聞き取りを原則とした．1つの姓に対して被験者から書き取り終了の合図が出るのを確認して，実験担当者は次の姓を回線に流す．これを4,000種の姓に対して繰り返した．被験者の聞き取り結果の記入方法として，分からない場合のみ空欄を認めた．試験に用いた姓は，日頃から良く耳にする姓と聞き慣れない珍しい姓が均等になるように実在頻度順位[9]が均一になるように選択した(星野他 1993)．試験に使用した姓4,000種の実在頻度順位，及び被験者の平均正解率を表1に示す．

表1から実在頻度順位が上位の姓ほど正解率が高いと言える．特に実在頻度順位5,000位以内の姓に対する平均正解率は93.8%，実在頻度順位5,000位から10,000以内の姓は93.7%と非常に高いが，実在頻度順位110,000位以降の姓に対する平均正解率は3割に満たない．

次に，人間同士の対話において正しく聞き取られることが多い姓，逆に聞き間違えられやすい姓の特徴をつかむために，被験者の聞き間違えに着目した．被験者の聞き間違えのうち，不正解であった姓の実在頻度順位 (X軸) に対し，間違えた先の実在頻度頻度 (Y軸) を図1にプロットした．図1から，プロットはグラフ下部に集中していることが分かる．すなわち被験者の聞き間違え先は，実在頻度順位上位に集中する傾向が見られる．分析したところ，不正解の

7 5名が20代～30代の男性，5名が30代女性
8 ナレータ業務を専門とする女性による録音を使用
9 日本全国，実在件数が多い順に並べた時の順位

表 1　聞き取りに使用した 4,000 種の個人姓と被験者の聞き取り精度

実在頻度順位	データ数 (件)	平均正解率 (%)
1 位 ～ 5,000 位	400	93.8
5,001 位 ～ 10,000 位	400	93.7
10,001 位 ～ 20,000 位	400	73.9
20,001 位 ～ 30,000 位	400	62.6
30,001 位 ～ 40,000 位	400	51.1
40,001 位 ～ 50,000 位	400	48.1
50,001 位 ～ 60,000 位	200	60.6
60,001 位 ～ 70,000 位	200	59.2
70,001 位 ～ 80,000 位	200	60.8
80,001 位 ～ 90,000 位	200	64.9
90,001 位 ～ 100,000 位	200	48.2
100,001 位 ～ 110,000 位	100	52.8
110,001 位 ～ 120,000 位	100	23.8
120,001 位 ～ 130,000 位	100	28.3
130,001 位 ～ 140,000 位	100	20.8
140,001 位 ～	200	14.0
計	4,000	60.8

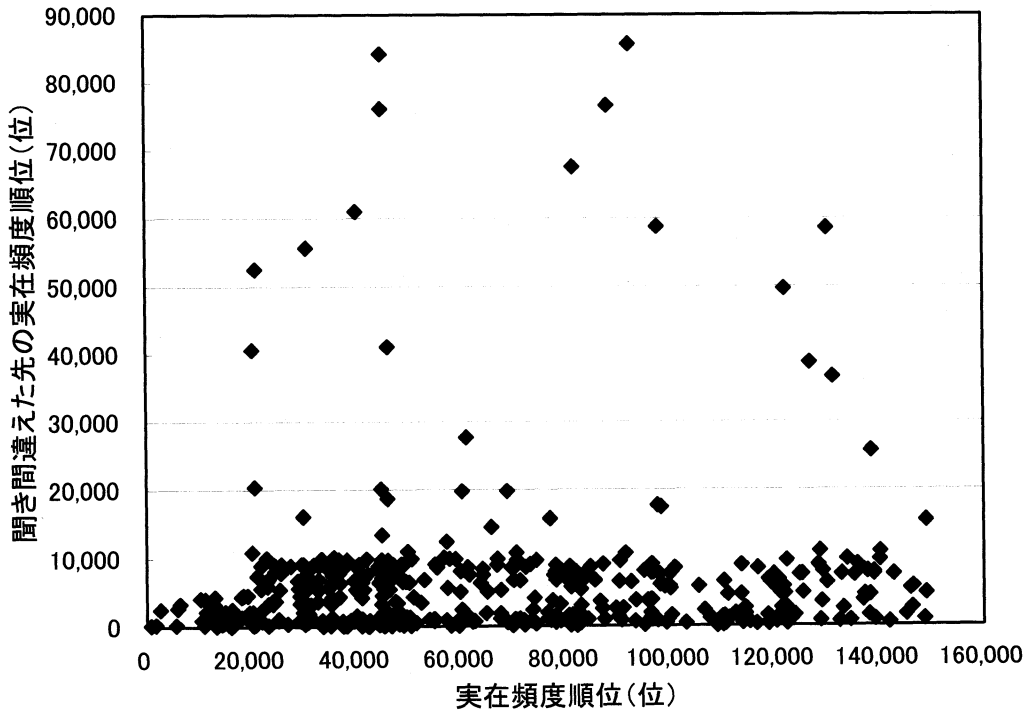

図 1　聞き間違えた姓の間違え先の実在頻度順位との関係

うち実在姓への聞き間違えは5,116件存在し，その約8割に該当する3,990件は実在頻度順位10,000位以内の姓に聞き間違えている．また，実在頻度順位20,000位以降の姓に間違えたものは全体の1割に満たない．更に5,116件のうち99.5%は，自分自身よりも頻度順位が上位の姓に聞き間違えている．

つまり人間は，個人姓の聞き取りにおいて，予め16万種全てを把握しているわけではなく，頻度順位上位の姓の中から正解を探し出そうとしていると言える．従って，聞き手は思い込んだ姓が発話された場合は正しく聞き取れるが，思い込み範囲外の姓に対しては，思い込んだ中で最も類似した姓に聞き間違える．この結果から人間同士の対話における思い込み戦略が確認できた．

3.2 思い込み戦略のインタフェースへの適用効果

3.1節の人間同士の対話の分析から，システムが誤認識結果を提示してしまうこと自体が利用者のストレスの要因なのではなく，人間の対話では起こりえないような聞き間違えが原因と考えられる．従って，人間が聞き間違えやすい対象を，システムが同じように間違えたとしても利用者は自然に受け止めると思われる．我々は，思い込みの仕組みを取り入れたシステムを構築するために，人間が聞き間違えないようなもののみを最初に認識対象とする．

思い込み戦略の適用は，正しい認識結果を利用者に提供するために対象語数を制限しなければならないという認識技術の限界に対する一時的な解決に繋がる．すなわち，大語彙全体を認識対象とする場合と比較して，一部しか認識対象としないため認識処理時間の短縮，及び認識精度の向上が見込める．

4 認識エンジンを用いた思い込み対象の分析

思い込み戦略において，利用者からアクセスされやすい対象を数多く思い込むほど，利用者に正解が提示できる可能性は高くなる．しかし，認識精度と網羅率はトレードオフの関係にあり，対象数の増加に伴い認識精度は低下する．本章では，市販の認識エンジンを利用した実験を通して，思い込み対象の選択方法について分析する．

4.1 認識精度と網羅率の関係

本節では個人姓に焦点を当て，認識エンジンNuance[10]を使用し，思い込み対象として選択すべき姓について分析した．利用者アンケートを通して，姓に対する網羅率と認識精度が利用者満足度に与える影響について述べる．

Nuanceには，予め姓の仮名表記リストを認識用文法として与える．被験者が姓を入力する

10 http://www.nuance.com

と，Nuanceは与えた認識用文法中の各姓に対して尤度を計算し，尤度の高い順に候補を提示する．提示候補数はデフォルト値の10とした[11]．すなわち最大で10候補が提示される．提示は画面に文字列で出力した．被験者は3.1節で聞き取りを依頼した10名である．最適な思い込み対象数を調べるために，実在頻度順位1位から対象数を変化させた個人姓認識用文法を15種用意した．15種の文法は含まれる個人姓の数のみが異なり，被験者に与える実験環境に差異はない．被験者に予め選択した個人姓400種の発話を依頼した．400種は，実社会と同じ実在頻度件数の分布を構成するために，各認識用文法の網羅率と一致するよう選択した．表2にNuanceに与えた認識用文法を構成する個人姓の実在頻度順位と網羅率(星野他 1993)，及び被験者に発話を依頼した400種の実在頻度順位の内訳と複雑度を示した．複雑度とは，聞き取り対象である姓の音韻上の複雑さの程度を表す情報量である．

姓Lにおける音韻列$W_1^n = w_1 \cdots w_n$の生成確率を$P(W_1^n)$とすれば，姓Lのエントロピーは，式(1)より計算できる(北 1999)．

$$H_0(L) = -\sum_{W_1^n} P(W_1^n) \log P(W_1^n) \tag{1}$$

複雑度とは一音韻あたりのエントロピーに相当し，式(2)より計算できる．各音韻の後には平均して$2^{H(L)}$種の音韻が後続可能であることを意味し，複雑度が大きいほど特定が困難なタスクである．

$$H(L) = -\sum_{W_1^n} \frac{1}{n} P(W_1^n) \log P(W_1^n) \tag{2}$$

網羅率は，日本の総人口1億2,000万人に対して該当する姓を持つ実在者数を計算した値である．表2より16万種類存在する個人姓のうち，実在頻度順位上位5,000種の姓で実在者数の約9割を網羅できることから，個人姓は実在頻度に大きな偏りがあることが分かる．また，表2に記述した各実在頻度順位の区切りは，被験者に発話を依頼した姓が認識用文法に含まれる割合を示すため，15種の認識用文法の構成数の区切りと一致させた．すなわち，認識用文法Aの中には，被験者の入力姓400種のうち72.3%にあたる289種が含まれ，この289種に対して尤度が計算されることになる．従って，残りの111種は認識用文法Aに含まれていないため結果に出力されることはない．

表3に，被験者の入力に対する平均認識精度を各文法毎に示した．表2及び表3より，認識対象数が多くなるほど複雑度は大きくなる．現在実用化されている住所確定インタフェースの初回の入力対象に設定されることが多い47都道府県名の平均複雑度は9.2，株価照会システムの入力対象に設定されている約700社の一部上場企業名の平均複雑度は10.1である．また，星

11 N-best値で設定する．

占いに用いられる12星座名の平均複雑度は8.1と小さい．これに対して，日本人の姓16万種の平均複雑度は16.7，名8万種の平均複雑度は14.9と大きい．現状の認識技術を用いてサービス提供が可能な認識対象数は，複雑度から考えると10程度，表2から10,000語程度が限界と考えられる．従って，姓や名は複雑度からも認識が非常に困難な対象であると言える．

最適な思い込み対象の選択方法を分析するために，認識エンジンの出力結果がインタフェースの第一応答として受け入れ可能な精度か否かを被験者に尋ねた結果を表3の最右欄に示した．90％以上の認識精度を持つ認識用文法C，D，Eに対しては，全被験者が音声対話インタフェースの第一応答として受け入れ可能と答えた．網羅率に関して分析すると，全被験者が受け入れ可能と判断した認識用文法C，D，Eの網羅率は90％以上である．網羅率が95％以上でも，認識精度が90％以下である認識用文法F以降に対しては，利用者満足度は獲得できていない．同様に認識用文法A，Bに関しては，認識精度が90％以上でも網羅率が低いため，思い込み範囲外の姓が発話される確率が高く正解が提示できない場合が多いことから，利用者のストレスに繋がると考えられる．

表 2　Nuanceに与えた認識用文法及び入力に使用した400種の個人姓

文法名	頻度1位からの選択数 (件)	網羅率 (%)	含まれる発話姓数 (件)	複雑度
A	1,000	72.3	289	8.2
B	3,000	86.2	345	8.9
C	5,000	90.6	362	9.3
D	8,000	94.1	376	9.4
E	10,000	95.6	382	9.8
F	15,000	97.1	388	10.6
G	20,000	98.0	392	11.2
H	30,000	98.9	395	11.8
I	40,000	99.4	397	12.6
J	50,000	99.6	398	13.1
K	60,000	99.7	398	13.9
L	70,000	99.7	398	14.7
M	80,000	99.9	399	15.1
N	90,000	99.9	399	15.5
O	100,000	99.9	399	15.8
P	160,000	100.0	400	16.7

4.2　人間と認識エンジンの思い込み結果の比較

本節では，人間同士の対話との比較を通して思い込み戦略の有効性を検証する．4.1節で全被験者が受け入れ可能と判断した中で，網羅率が最大の文法EをNuanceに思い込み対象として与える．

表 3　認識用文法毎の認識結果及び被験者による受け入れ可否評価

文法名	網羅率 (%)	平均認識精度 (%)	受入れ可人数 (10 名中)
A	72.3	94.8	6
B	86.2	95.9	7
C	90.6	94.1	10
D	94.1	91.2	10
E	95.6	90.8	10
F	97.1	80.3	8
G	98.0	76.5	8
H	98.9	66.3	7
I	99.4	67.2	4
J	99.6	62.1	4
K	99.7	57.4	2
L	99.7	42.1	0
M	99.9	44.7	0
N	99.9	48.3	0
O	99.9	44.4	0
P	100.0	42.5	0

入力には，3.1 節で個人姓の聞き取りに用いたナレータの録音音声 4,000 種のうち，実在頻度順位 10,000 位以内の姓 400 種を用いた．入力 400 種に対する認識結果を表 4 に示す．認識エンジンは頻度順位上位の姓 10,000 件を思い込み対象とした場合，3.1 節の試験における被験者の聞き取りとほぼ同じ精度を持った認識結果を返すことができる．

表 4　認識エンジンの出力結果と人間の聞き取り結果の比較

実在頻度順位	試験件数 (件)	平均認識精度 (%)	
		認識エンジン	被験者
1 位 ～ 5,000 位	200	93.9	93.8
5,001 位 ～ 10,000 位	200	91.2	93.7
計	400	92.6	93.8

次に，思い込み対象外の姓が発話された場合の認識エンジンの結果について考察する．今度は，3.1 節のナレータ録音音声 4,000 種のうち，実在頻度順位 10,000 位以降の 3,600 種を入力として Nuance に与えた．Nuance は与えられた思い込み対象文法 E に含まれる姓に対して尤度計算するため，入力姓 3,600 種に対する結果は全て誤認識となる．表 5 は，3.1 節で被験者が実在姓に聞き間違えた 5,116 件のうち，被験者の間違えた姓と Nuance が尤度 1 位を算出した誤認識結果が一致した割合を示したものである．文法 E を思い込み対象とした場合，人間の聞き取りとほぼ同じ精度を持ち，思い込み対象外の発話に関しては聞き間違える先もほぼ一致する．従って，認識結果が誤認識であっても，人間同士の対話でも起こりえる間違え方であることか

表 5　認識エンジンと人間の聞き間違え一致度

実在頻度順位	聞き間違え一致度 (%)
10,001 位 ～ 20,000 位	87.3
20,001 位 ～ 30,000 位	87.9
30,001 位 ～ 40,000 位	86.5
40,001 位 ～ 50,000 位	89.1
50,001 位 ～ 60,000 位	91.1
60,001 位 ～ 70,000 位	87.3
70,001 位 ～ 80,000 位	91.4
80,001 位 ～ 90,000 位	90.9
90,001 位 ～ 100,000 位	88.3
100,001 位 ～ 110,000 位	91.8
110,001 位 ～ 120,000 位	90.9
120,001 位 ～ 130,000 位	90.8
130,001 位 ～ 140,000 位	88.1
140,001 位 ～	88.8
平均	73.4

ら利用者はストレスを感じないと考える．

5　大語彙インタフェースの実装

我々は，大語彙インタフェースの実現を目指している．検討の第一歩として，利用者にストレスを与えない応答を返すために，思い込み戦略を取り入れた聞き取り手法を提案する．これを「思い込み応答」と呼ぶこととする．一般に人間同士の対話における “応答” は，相手の発話に対して確認を伴うとは限らないが，思い込み応答の “応答” は，聞き取った結果を話し手に提示，確認する行為を表すことにする．思い込み応答とは，思い込んだ範囲内から聞き取った結果を探し出し，提示する聞き取り確認手法である．

しかし，音声対話インタフェースは利用者の入力を確定することが最終目的であり，思い込みが外れた場合の対応も考える必要がある．本章は，思い込み応答にて正解を提示できない場合に利用者にストレスを与えない仕組みを提案し，この仕組みと思い込み応答を組み合わせたインタフェースを実装する．

5.1　思い込みが外れた場合の人間の反応

思い込み範囲外の対象が発話された際の人間の対話を分析するために，3.1節の聞き取り試験における被験者 10 名に実在頻度順位 100,000 位以降の希少姓 100 種の聞き取りを再度依頼した．発話は 3.1節で 4,000 種の個人姓を発話したナレータに再度依頼した．被験者には聞き取れるまで自由な質問を許し，発話者には被験者からの質問に対して Yes 又は No の正誤応答と姓

の再発話のみを応答として許容した．10名の被験者による計1,000件の第一応答は，701件の再発話要求(「もう一度お願いします」)と，299件の聞き取り結果の正誤確認質問(「～さんで正しいですか？」)の2種に分類された．後者の299件中，正解は24件のみで，残りの275件は誤認識であった．発話者からNoと返された275件に対して，被験者全員が再発話を要求した．この結果を利用して，我々は，利用者の入力に対して確度の高い認識結果が得られなかった場合は思い込みが外れたと判断し，利用者に再入力を要求するインタフェースを構築する．

5.2　思い込み応答のインタフェースへの組み込み

思い込みが外れた場合に利用者にストレスを与えない対応をするために，思い込み対象姓に対する認識結果を応答として利用者に提示する時間及びそれに対する利用者の回答時間内に，思い込み範囲外の姓に対して並行認識することを提案する．これにより，思い込み範囲外の発話に対して裏認識結果から正解が出力される可能性が生じる．

実用インタフェースではリアルタイム性と精度が要求される．第一応答で正解が提示できない場合，第二応答提示時には裏認識を終了させ正解が提示できることが好ましい．精度の観点から，思い込み範囲外の残りも大語彙であり，裏認識で一度に認識対象としても精度は期待できない．そこで思い込み範囲外を精度が期待できる対象数毎に分割し，各文法に対する裏認識を提案する．

我々は，以下のような思い込み適用フローを考案し，Nuanceを利用して実装した．4章の分析に基づき，頻度順位上位10,000件の姓からなる文法Eを思い込み対象としてNuanceに与えた．利用者が入力すると，文法Eに含まれる姓に対して尤度計算する．第一応答では，閾値以上の尤度を持つ候補を利用者に提示し正誤確認を行うこととする．閾値以上の尤度を持つ候補が出力されない場合は，思い込みが外れたと判断し姓の再入力を要求する．システムは，思い込みの結果を応答として利用者に提示する時間，及びそれに対する利用者の回答時間を利用して裏認識を行う．複雑度を基に考えると，裏認識における各文法の語数は10,000件程度が限界であると考えた．現状Nuanceの認識処理時間は，N-best = 10の場合，10,000語の認識用文法に対して0.58秒要する[12]．実装フローにおいて，閾値以上の尤度を持つ候補を提示するガイダンス「入れていただいたお名前は～さんですか？」の出力に3.99秒，閾値以上の尤度を持つ候補が出力されない場合，姓の再入力を要求ガイダンス「もう一度お名前をお願いします」の出力に3.47秒を要する．それに対して利用者は，Yes又はNoの正誤回答に平均1.12秒，再入力に平均1.98秒要する．従って10,000語ずつに15分割すると，第二応答提示時までに15回の裏認識処理は終了可能なのでリアルタイム性も提供できる．裏認識では，最初の利用者の入力をシステム内部に録音したものを入力として用いる．フロー概要を図2に示す．第二応答は，思い込み範囲外の分割した各文法の裏認識結果を算出尤度順に並べ，最も尤度の高い姓を利用

12 10,000回の認識試験の平均CPU使用時間

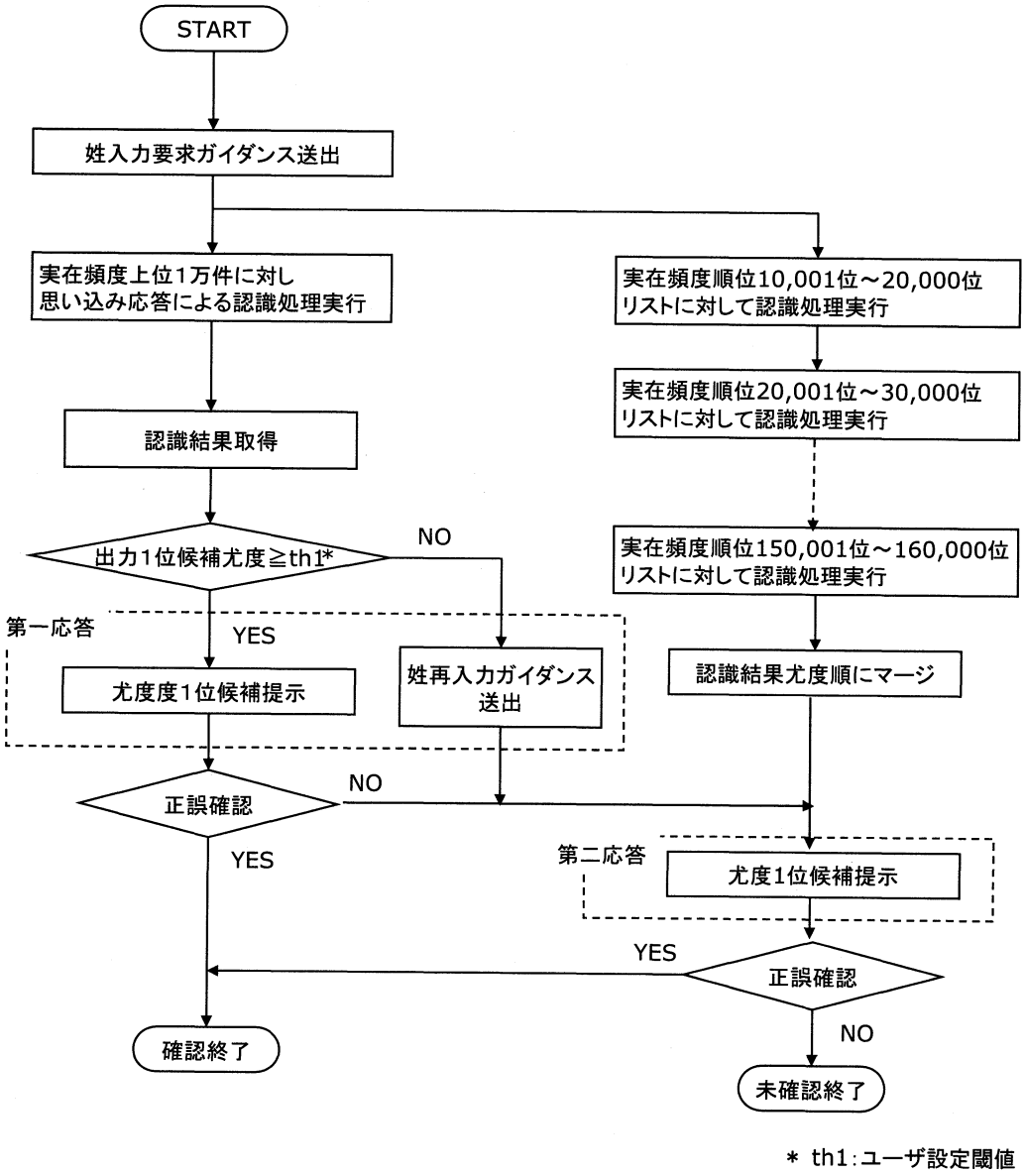

図 2　思い込み適用フロー

者に提示する．思い込みによる第一応答で正解候補を利用者に提示できた場合は，裏認識は不要であったことになる．裏認識処理の併用により，思い込みが外れても第二応答でリアルタイムに正解を提示できる可能性が大いにあると考えられる．

5.3　評価実験結果

本節では，5.2節で述べた思い込みを適用したフローを実装し，思い込み応答の有効性を検証する．評価のために，思い込みを適用しないフローも実装した．このフローでは16万種の姓を一度に認識用文法としてNuanceに与える．閾値以上の尤度を持つ候補が出力された場合は利用者に提示し正誤確認を行う．閾値以上の候補が提示されない場合，又は提示が否定された場合は，利用者に再入力を要求する．再入力された姓に対して，16万種を対象として再認識処理を行い，第二応答は尤度1位の候補を無条件に提示する．提示が誤認識の場合は未確定のまま終了する．思い込み適用フローは，思い込み応答及び裏認識で用いる文法は10,000語の小語彙であるため，未適用フローに比べ精度が期待できる．更に，第一応答は誤認識であっても人間同士の対話でも起こりえるような聞き間違えであることから，利用者はストレスを感じないと考える．実験において，システムの正誤確認に対する利用者の回答は100%の精度で獲得可能であるものとする．

新たな被験者20名[13]に，思い込み未適用フロー及び適用フローの2種に対して5.1節の希少姓100種の入力を依頼した．被験者に，両フローの第一応答，及び第二応答が音声対話インタフェースの応答として受け入れ可能か否かの判断を依頼した．第一応答及び第二応答に対する評価を独立に行うため，被験者はシステムからの第一応答提示直後に第一応答を評価し，第二応答提示直後に第二応答の評価を入力1件毎に行う．表6に両フローの第一応答，第二応答の応答内容の内訳，及び被験者の受け入れ可否評価を応答種別毎に示した．

表6　思い込み応答の有効性検証

	第一応答			第二応答	
	正解 (件)	再入力 (件)	誤提示 (件)	正解 (件)	誤提示 (件)
思い込み未適用	175	490	1,335	202	1,623
(受入れ可能応答)	(175)	(234)	(0)	(202)	(121)
思い込み適用	0	1,053	947	1,438	562
(受入れ可能応答)	(0)	(426)	(670)	(1,438)	(307)

思い込み未適用フローは，入力計2,000件のうち，第一応答で正解を提示できたものが175件，再入力要求が490件，残り1,335件は誤った候補の提示であった．この第一応答に対して被

13 20～30代の男女各10名ずつ

験者は，正解提示の 175 件及び再入力要求の一部 234 件の計 409 件 (全体の約 20 %) をインタフェースの第一応答として受け入れ可能と評価した．誤提示に対しては全て受け入れ不可と評価した．再入力後の第二応答で正解を提示できたものは 202 件のみであり，全体の約 80 %にあたる 1,623 件は誤提示となり，未確定のまま終了した．第二応答の評価対象 1,825 件のうち，正解提示の 202 件及び誤提示の一部 121 件の計 323 件 (全体の約 18 %) を受け入れ可能と評価された．ここで，第二応答で受け入れ可能と評価された誤提示の一部 121 件の第一応答は，全て再入力を要求している．このことから，第一応答と第二応答で誤提示が繰り返された場合，利用者はストレスを感じると言える．

一方，思い込み適用フローでは，各被験者の入力 100 件はいずれも思い込み範囲外の姓であるため，第一応答で正解が出力されることはない．第一応答は 1,053 件が再入力要求，947 件が誤提示であり，再入力 1,053 件中の約 40 %にあたる 426 件，及び誤提示 947 件中の約 70 %にあたる 670 件の計 1,096 件 (全体の 54 %) が第一応答として受け入れ可能と評価された．第二応答では正解を提示できたものが 1,438 件，全体の約 28 %にあたる 562 件が誤提示となり未確定のまま終了した．第二応答に対して被験者は，正解提示の 1,438 件及び誤提示の一部 307 件の計 1,745 件 (全体の約 87 %) を受け入れ可能と評価した．思い込み未適用の場合と同様に，誤提示で受け入れ可能と評価された 307 件の第一応答は，全て再入力を要求している．

思い込みを適用しない場合，約 8 割の入力が未確定のまま終了し，第一応答，第二応答ともに受入れ不可という判断が約 8 割を占める．第一応答に着目すると，受け入れ可能と判断されたのは正解提示と再入力を要求した応答の一部のみであり，誤提示に対しては全て受け入れ不可と評価された．これに対して思い込み適用フローでは約 8 割が確定終了した．思い込みを適用した第一応答に着目すると，誤提示対話の約 7 割，再入力要求を合わせると全体の半数以上が受け入れ可能と判断された．第二応答に対しては約 9 割が受け入れ可能と評価された．

思い込みを適用した応答は，誤認識であっても利用者には受け入れられることが分かる．このことから，誤認識を生じること自体が必ずしも利用者のストレスの要因ではないことが確認できた．思い込み適用フローにおいて未確定に終った 562 件の存在が，第二応答に対して受け入れ不可と判断された約 1 割の応答の存在に繋がると考えられる．

6　まとめ，及び今後の課題

我々は，思い込み戦略，及びそれを取り入れた思い込み応答を提案することで，大語彙を入力対象とした実用インタフェースにおいて利用者にストレスを与えない応答を返す仕組みを実現した．また，思い込み応答による聞き取りの精度は人間同士の対話とほぼ同じであり，聞き間違え先も人間の聞き取りとほぼ一致することを確認した．5.3 節の実験結果から，思い込みが外れた場合，裏認識処理のみでは第二応答で正解を提示できないことがある．更に実験結果から，第一応答と第二応答で誤提示が繰り返された場合は利用者に受け入れられないことが分

かった．1章で述べたように，正誤確認と再入力の繰り返しは利用者のストレスに繋がる．我々は，思い込みの結果，利用者に正解を提示できなかった場合，利用者に別の質問をすることで迅速に正解を絞り込む方法を検討している．すなわち，絞り込みのための質問を利用者にストレスを与えない順序で，かつ利用者からの質問に対する回答も誤認識である場合を考慮して，質問を組み立てる必要がある．

個人姓名に関しては，絞り込みに有効な質問の選定が大きな課題であると考える．「サトウ様ですか？」「カトウ様ですか？」「アトウ様ですか？」のような候補の提示と再入力の繰り返し，更に再入力後も誤認識を繰り返すことは利用者にストレスを与えるのは明らかである．人間は，相手の苗字や名前を聞き取れなかったと感じた時，漢字表記や頭文字を尋ねたり，曖昧な部分について問いかけをしながら正しい姓名を導きだそうとする．我々は，人間同士の対話の分析を進めることで，姓名確定のための対話制御方式の確立を目指す．

一方，住所に関しては，サービスのアクセス頻度の偏りやコールセンターへの発信番号から思い込み対象を効果的に選択することで，利用者に都道府県名からの入力を強制する必要がなくなる．2.2節で述べたコールセンターのオペレータは，町村名が聞き取れない場合，再入力要求ではなく上位階層である都道府県名，或いは市区郡名を尋ねる傾向が見受けられた．この分析から住所に関しては，思い込みが外れた場合，階層構造を利用して対象数の少ない上位階層を確定する方向へ対話を進めた際の有効性について検証を進めている．その他の大語彙として，全国1,100万種の登録がある企業名の確定[14]や年間約9万回主催されるコンサートの特定[15]などへの適用も検討している．

これまで本稿は，音声入力の応答を例に挙げ議論を進めてきたが，思い込み応答は，音声入力に限らず，大量の検索空間から利用者が必要とする情報を高速かつ高性能に検索する手段として役立つと考える．検索空間が広範囲に及ぶ情報検索の分野では，検索キーに対して数多くの検索結果が取得できてしまい，情報を絞り込む手段がないのが現状である．利用者毎の行動履歴やアクセス履歴を基に，思い込み戦略を利用者の検索趣向を導き出す手段に適用することで，同じ検索キーが入力された場合でも個々人適応型の情報提供が可能になると考えられる．

今後，思い込み戦略の他分野への適用を検討すると同時に，思い込みにより正解を提示できない場合に，迅速に正解を導くための対話制御方法の検討を続けていきたい．

謝辞

本論文に関してご指導頂きました (株) NTT アドバンステクノロジー東田正信氏，NTT コミュニケーション科学基礎研究所堂坂浩二氏，本研究の機会を与えて下さいました (株) NTT データ技術開発本部長松本隆明氏，的確で有益なコメントを下さいました査読者の方に深く感謝致します．

14 職業別電話帳「タウンページ」掲載数
15 定期刊行雑誌「ぴあ」(ぴあ株式会社)12ヶ月分の集計

参考文献

赤堀一郎，加藤利文，北岡教英 (1995). “地名認識システムとその応用.” 情報処理学会研究会報告, **SLP-7-9**.

荒井和博，吉岡理，嵯峨山茂樹，山田智一，野田喜昭，井本貴之，菅村昇 (2000). “音声認識機能を持つ住所入力システム.” 情報処理学会研究会報告, **SLP-5-10**.

Chu-Carroll, J. (2000). “MIMIC: an adaptive mixed initiative spoken dialogue system for information queries.” In *Proceedings of the ANLP-2000*, pp. 97–104.

堂坂浩二，安田宜仁，相川清明 (2002). “システム知識制限下での効率的音声対話制御方法.” 言語処理学会, **9** (1), pp. 43–63.

Ferguson, G. and Allen, J. F. (1998). “TRIPS: An Integrated Intelligent Problem-solving Assistant.” In *Proceedings of the ANLP-2000*, pp. 567–572.

Godden, D.,Brill, E.,Glass, J.,Pao, C.,Phillips, M.,Porifroni, J.,Sneff, S.,and Zue, V. (1994). “GALAXY: A human-language interface to on-line travel information.” In *Proceedings of the ICSLP-94*, pp. 707–710.

Godden, D.,Meng, H.,Polifromi, J.,Seneff, S.,and Busayapongchai, S. (1996). “A frame-based dialogue manager for spoken language applications.” In *Proceedings of the ICSLP-96*, pp. 701–704.

星野裕，加藤博子，永田健児 (1993). 30 万人読み方書き方辞典. 日外アソシエーツ株式会社.

自治省 (1998). 国土行政区画総覧. 国土地理協会.

亀田弘之，藤崎博也 (1997). “情報検索における音声・言語情報処理.” 情報処理学会研究会報告, **SLP-25-14**.

菊地英明，白井克彦 (2000). “対話効率の向上を目的とした音声対話制御のモデル化.” ヒューマンインタフェース学会誌, **2** (2), pp. 145–152.

伊藤亮介，駒谷和範，河原達也 (2002). “機器操作マニュアルの知識と構造を利用した音声対話ヘルプシステム.” 情報処理学会論文誌, **43** (7), pp. 2147–2154.

北研二 (1999). 確率的言語モデル. 東京大学出版会.

村上仁一，嵯峨山茂樹 (1991). “自由発話音声認識における音響的および言語的な問題点の検討.” 日本音響学会音声研究会資料, **SL91-100**, pp. 71–78.

Nakano, M.,Dohsaka, K.,Miyazaki, N.,Hirasawa, J.,Tamoto, M.,Kawamori, M.,Sugiyama, A.,and Kawabata, T. (1999). “Handling rich turn-taking in spoken dialogue system-s.” In *Proceedings of the Eurospeech '99*, pp. 1167–1170.

Nielsen, P. B. and Baekgaard, A. (1992). “Experience with a dialogue description formalism for realistic applications.” In *Proceedings of the ICSLP-92*, pp. 719–722.

新美康永，小林豊 (1995). “音声認識の誤りを考慮した対話制御方式のモデル化.” 情報処理学

会研究会報告, **SLP-5-7**.

新美康永 (1998). "音声対話システムの対話制御." 日本音響学会誌, **54** (11), pp. 791–796.

Niimi, Y.,Nishimoto, N.,and Kobayashi, Y. (1997). "Analysis of interactive strategy to recover from misrecognition of utterances including multiple information items." In *Proceedings of the Eurospeech '97*, pp. 2251–2254.

Niimi, Y.,Takigawa, N.,and Nishimoto, T. (1995). "Modeling dialogue strategies to resolve speech recognition errors." In *Proceedings of the Eurospeech '95*, pp. 534–537.

古井貞熙 (1998). 音響・電子工学. 近代科学社.

吉岡理, 荒井和博, 菅村昇, 嵯峨山茂樹 (1997). "音声認識機能を含むマルチモーダルインタフェースをもつ住所入力システムの開発と評価." 電子情報通信学会論文誌, **J80-D-II**, pp. 1007–1015.

Zeigler, B. and Mazor, B. (1995). "Dialogue design for a speech-interactive automation system." In *Proceedings of the Eurospeech '95*, pp. 113–116.

Zue, V.,Seneff, S.,Glass, J.,Goddeau, D.,Goddin, D.,Pao, C.,Phillips, M.,and Porifroni, J. (1994). "PEGASUS: A spoken dialogue interface for on-line air travel planning." *Speech Commnications*, **5**, pp. 331–340.

略歴

大森 久美子: 平成 8 年慶應義塾大学大学院理工学研究科計算機科学専攻終了. 同年, 日本電信電話 (株) 情報通信研究所入社. 平成 10 年より同社情報流通プラットフォーム研究所にて音声対話処理の研究に従事, 現在 (株) NTT データ技術開発本部にて音声対話制御手法の研究, アプリケーション開発に従事. 平成 15 年 4 月より, 慶應義塾大学大学院理工学研究科開放環境科学専攻計算機科学専修後期博士課程在籍. 自然言語処理, 音声言語理解に興味を持つ. 情報処理学会, 言語処理学会, 電子情報通信学会各会員.

斎藤 博昭: 慶應義塾大学工学部数理工学科卒業. 現在同大理工学部情報工学科専任講師. 工学博士. 自然言語処理, 音声言語理解などに興味を持つ. 情報処理学会, 言語処理学会, 日本音響学会, 電子情報通信学会, ACL 各会員.

(2002 年 12 月 25 日 受付)
(2003 年 3 月 24 日 再受付)
(2003 年 5 月 19 日 再々受付)
(2003 年 5 月 21 日 採録)